Overestimating Forecast Skill Through Improper Application of Verification Metrics: Simpson's Paradox in Meteorology

Thomas M. Hamill

NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado

Josip Juras

Geophysical Institute, Faculty of Science, University of Zagreb

Zagreb, Croatia

(Revised)

Submitted to *Monthly Weather Review*

17 August 2005

Corresponding author address:

Dr. Thomas M. Hamill
NOAA-CIRES Climate Diagnostics Center
R/CDC 1, 325 Broadway
Boulder, CO 80301 USA

e-mail: tom.hamill@noaa.gov
phone: 1 (303) 497-3060

ABSTRACT

It is common practice to summarize the skill of weather forecasts using an agglomeration of samples spanning many locations and dates. In calculating many of these verification metrics, there is an implicit assumption that the climatological frequency of event occurrence is fixed for all samples. If the event frequency actually varies among the samples, then the scores may report fictitiously high skill. This is an example of the previously described statistical conundrum known as "Simpson's Paradox." Many common deterministic verification metrics such as threat scores are subject to overestimation of skill, and probabilistic forecast metrics such as the Brier skill score and relative operating characteristic are also affected. Demonstrations of the false skill are provided, and guidelines are suggested for how to adapt these diagnostics to avoid this problem.

1. **Introduction**

This article will demonstrate that many commonly used weather forecast verification metrics are capable of reporting positive forecast skill when none truly exists or reporting more skill than the forecast truly has. Depending on the metric and the event being verified, this effect can be large or small.

This effect is has been described in the conventional statistics literature, where it is known as "Simpson's Paradox" (Simpson 1951, Appleton et al. 1996, Malinas and Bigelow 2004). Tables 1-3 provide a common illustration of this paradox, using hypothetical rates of admission into graduate-school physical science programs. Assume that there are two universities of higher education in a state system, and Tables 1 and 2 provide the admittance data for each university. At both universities, the men are admitted more frequently than the woman. However, Table 3 composites the data from the two universities, and taken together, men appear to be admitted less frequently than women. Would men have reason for a gender discrimination lawsuit against the state? Clearly not; the problem is that the first grad school was not very selective in their admissions, the second was very selective, and this extra factor, a "lurking variable" in statistical parlance, confounded the interpretation when lumping the two together.

A similar problem occurs in meteorological forecast verification, but it is not widely appreciated. We have encountered circumstances where we have diagnosed large positive forecast skill when intuition suggested that little or no skill would exist. For example, the first author used a common probabilistic forecast verification metric, the relative operating characteristic, in a comparison of ensemble forecast methods (Hamill et al. 2000b, Fig. 13). The author reported a relative operating characteristic curve for wind speed forecasts at 5 days lead that indicated a highly skillful forecast, different than experience would suggest for this lead time. The second author discussed the overestimation of forecast skill (Juras 2000) in a comment on a Buizza et al. (1999) article. It was indicated that the chosen metrics might report false skill if climatological event frequencies vary within the verification area. This issue has also been raised in Mason (1989) and less directly in other meteorological publications, including Buizza (2001; p. 2335), Stefanova and Krishnamurti (2002, p. 543), Atger (2003), Glahn (2004; p. 770), and Göber et al. (2004). Still, there are hundreds of published articles that should have but did not factor in this effect, including two by the lead author (Hamill 1999, Hamill et al. 2000b). Clearly, the problem is not appreciated as widely as it should be.

In this article we will examine three common skill metrics, the Brier skill score (Wilks 1995), the relative operating characteristic (Swets 1973, Harvey et al. 1992), and the equitable threat score (Schaefer 1990). All are capable of reporting positive forecast skill when none is present. Many other metrics such as the ranked probability skill score (Wilks 1995, Epstein 1969, Murphy 1971), economic value diagrams (Richardson 2000, Palmer et al. 2000, Richardson 2001b, Zhu et al. 2002, and Buizza et al. 2003), and other contingency-table based threat scores will not be discussed but are subject to the same problem.

Section 2 will provide a brief review of the three chosen verification metrics, as well as descriptions of how they are computed. Section 3 follows with a very simple example of false skill and an explanation of why it occurs. Section 4 shows that the false value may or may not be reported with real meteorological data, depending on what event is being considered. Section 5 demonstrates how large the effect can be for a common verification problem, the threat scores of short-range precipitation forecasts. Section 6 concludes with a discussion of the implications and how to adapt verification strategies to minimize or avoid this problem.

2. **Computation of common verification metrics**

Below, we review three general verification metrics, the Brier skill score, relative operating characteristic, and the equitable threat score. After the review, we describe how each of these metrics can be calculated in several different ways.

The long-used Brier score (Brier 1950) is a measure of the mean-square error of probability forecasts for a dichotomous (two-category) event, such as the occurrence/non-occurrence of precipitation. A review is provided in Wilks (1995), and references therein provide further background. The Brier score is often hard to interpret; is a Brier score of 0.06 good or bad? Consequently, the Brier score is often converted to a skill score, its value normalized by the Brier score of a reference forecast such as climatology or persistence (ibid). A Brier skill score (BSS) of 1.0 indicates a perfect probability forecast, while a BSS of 0.0 should indicate the skill of the reference forecast (see Mason 2004 for further discussion of whether a BSS of 0.0 indicates no skill).

The relative operating characteristic (ROC) has gained widespread acceptance in the past few years as a metric for ensemble forecast verification. The ROC has been used for decades in engineering, biomedical, and psychology applications; see an overview in Swets (1973). Its application in meteorology was proposed in Mason (1982), Stanski et al. (1989), and Harvey et al. (1992). In the Hamill et al. (2000a) summary of an ensemble workshop, it was recommended by the ensemble verification community as a standard metric, and the ROC was recently made part of the World Meteorological Organization's (WMO) standard

(WMO, 1992). Characteristics of the ROC have been discussed in Buizza et al. (1998), Mason and Graham (1999, 2002), Juras (2000), Wilson (2000), Buizza et al. (2000ab), Wilks (2001), Kheshgi and White (2001), Kharin and Zwiers (2003), and Marzban (2004). The technique has been used to diagnose forecast accuracy in, for example, Buizza and Palmer (1998), Buizza et al. (1999), Hamill et al. (2000b), Palmer et al. (2000), Richardson (2000, 2001ab), Wandishin et al. (2001), Ebert (2001), Mullen and Buizza (2001, 2002), Bright and Mullen (2002), Yang and Arritt (2002), Legg and Mylne (2004), Zhu et al. (2002), Toth et al. (2003), and Gallus and Segal (2004). Harvey et al. (1992) provide a thorough review of the concepts underlying the ROC.

The equitable threat score (ETS) provides one of many ways of summarizing the ability of a deterministic forecast to correctly forecast a dichotomous event. The ETS will produce a score of 1.0 for a perfect forecast, and random forecasts should be assigned a value of 0.0. The ETS is commonly used to evaluate the skill of forecasts, especially precipitation. See, for example, Rogers et al. (1995, 1996), Hamill (1999), Bayler et al. (2000), Stensrud et al. (2000), Xu et al. (2001), Ebert (2001), Gallus and Segal (2001), Chien et al. (2002), and Accadia et al. (2003).

The method for computing these metrics is now discussed, starting with the probabilistic metrics. The BSS and ROC will be generated from ensemble forecasts, though they can be generated from any probabilistic forecast.

Start by defining a dichotomous event of interest, such as occurrence/non-occurrence of precipitation, or temperature above or below a threshold. Let $\mathbf{X}_e(j,k) = [X_1(j,k), \ldots, X_n(j,k)]$ be an n-member ensemble forecast of the relevant scalar variable (again, precipitation or temperature) for the jth of m locations and the kth of r case days.

The ensemble at that day and location is first sorted from lowest to highest. This sorted ensemble is then converted into an *n*-member binary forecast $\mathbf{I}_e(j,k) = [I_1(j,k), \ldots , I_n\ (j,k)]$ indicating whether the event was forecast (=1) or not forecast (=0) in each member. The observed weather is also converted to binary, denoted by $I_o(j,k)$.

a. Brier skill scores

Assuming that each member forecast is equally likely, a forecast probability $p_f(j,k)$ is calculated from the dichotomized ensemble:

$$p_f(j,k) = \frac{\sum_{i=1}^{n} I_i(j,k)}{n} \qquad . \tag{1}$$

The Brier score of the forecast BS_f is calculated as

$$BS_f = \sum_{k=1}^{r} \sum_{j=1}^{m} \left(p_f(j,k) - I_o(j,k)\right)^2 \qquad . \tag{2}$$

A Brier skill score (BSS) is calculated as

$$\text{BSS} = 1.0 - BS_f / BS_c \ , \tag{3}$$

where BS_c is the Brier score of the reference probability forecast, commonly the probability of event occurrence from climatology.

An ambiguity and potential source of false skill may be traced to the method for calculating BS_c, if an appropriate long-term climatology is not available One method would be to generate a sample climatological probability $p_c(j)$ of event occurrence unique to each location of the *m* locations in the domain,

$$p_c(j) = \frac{\sum_{k=1}^{r} I_o(j,k)}{r} , \tag{4}$$

in which case BS_c would be

$$BS_c = \sum_{k=1}^{r}\sum_{j=1}^{m}\left(p_c(j) - I_o(j,k)\right)^2 \quad . \tag{5}$$

Another way would be to calculate a climatology p_c averaged over all locations

$$p_c = \frac{\sum_{k=1}^{r}\sum_{j=1}^{m} I_o(j,k)}{r \cdot m} \quad , \tag{6}$$

and let

$$BS_c = \sum_{k=1}^{r}\sum_{j=1}^{m}\left(p_c - I_o(j,k)\right)^2 \quad . \tag{7}$$

Differences in the calculation from using (4) – (5) instead of (6) – (7) will be illustrated in sections 3 and 4.

b. ROC diagrams

Calculation of the ROC starts with the population of 2x2 contingency tables, with separate contingency tables tallied for each sorted ensemble member and location. The contingency table for the *j*th location and *i*th sorted ensemble member has four elements: Γ_i(j) = [a_i(j), b_i(j), c_i(j), d_i(j)], indicating the relative fraction of hits, misses, false alarms, and correct rejections (Table 4). The contingency table is populated using data over all *r* case days, and then each is normalized so the sum of the elements is 1.0.

The hit rate (*HR*) for the *i*th sorted forecast and *j*th location is defined as

$$HR_i(j) = \frac{a_i(j)}{a_i(j) + b_i(j)} . \tag{8}$$

Similarly, the false alarm rate is defined as

$$FAR_i(j) = \frac{c_i(j)}{c_i(j) + d_i(j)}. \tag{9}$$

The ROC for the *j*th of *m* locations is a plot of *HR*i (*j*) (ordinate) vs. *FAR*i (*j*) (abscissa), *i* = 1, … , *n*. A ROC curve that lies along the diagonal *HR*=*FAR* line indicates no skill; a curve that sweeps out maximal area, as far toward the upper left corner as possible, indicates maximal skill. The ROC is commonly summarized through the integrated area under the ROC curve, or AUC. A perfect forecast has an AUC of 1.0, and climatology an AUC of 0.5.

It has often been judged to be more convenient to examine one rather than *m* different ROC curves. Hence, a single ROC is commonly generated from contingency tables averaged over all locations, i.e., $\Gamma_{\mathrm{i}} = \left(\overline{a}_{\mathrm{i}}, \overline{b}_{\mathrm{i}}, \overline{c}_{\mathrm{i}}, \overline{d}_{\mathrm{i}}\right)$ where $\overline{a}_{\mathrm{i}} = \sum_{j=1}^{m} a_{\mathrm{i}}(j) / m$, and $\overline{b}_{\mathrm{i}}, \overline{c}_{\mathrm{i}}$, and $\overline{d}_{\mathrm{i}}$ are similarly defined. Then

$$HR_i = \frac{\overline{a}_i}{\overline{a}_i + \overline{b}_i} \tag{10}$$

and

$$FAR_i = \frac{\overline{c}_i}{\overline{c}_i + \overline{d}_i} \tag{11}$$

c. Equitable threat score

Assume now that we have a deterministic forecast rather than an ensemble. With sufficient sample size, the ETS could be calculated for each *j* of the *m* locations using

Table 4 (but dropping the $_i$ subscript denoting the ensemble member number). The equation for the *ETS* is

$$ETS(j) = \frac{a(j) - a_r(j)}{a(j) + b(j) + c(j) - a_r(j)}, \tag{12}$$

where $a_r(j)$ is the expected fraction of correct forecasts for a random forecast

$$a_r(j) = \frac{\big(a(j) + c(j)\big)\big(a(j) + b(j)\big)}{a(j) + b(j) + c(j) + d(j)}. \tag{13}$$

Commonly because of small sample size, the ETS is calculated using contingency tables summed over all the grid points. Let $\overline{a} = \sum_{j=1}^{m} a(j) / m$, and define $\overline{b}, \overline{c}$, and $\overline{d}$ similarly. Then an ETS that presumably represents the domain-averaged skill is calculated from

$$ETS = \frac{\overline{a} - \overline{a}_r}{\overline{a} + \overline{b} + \overline{c} - \overline{a}_r}, \tag{14}$$

where

$$\overline{a}_r = \frac{\left(\overline{a} + \overline{b}\right)\left(\overline{a} + \overline{c}\right)}{\overline{a} + \overline{b} + \overline{c} + \overline{d}}. \tag{15}$$

3. **An example of false skill: synthetic data at two independent locations**

Suppose a hypothetical planet consists of one big ocean and two small, isolated islands, and suppose weather forecasting is utterly impossible on this planet; the best one can do is to forecast the climatological probability distribution appropriate to each island. To simulate this, assume that at island 1, the daily maximum temperature was randomly sampled from its fixed climatological distribution ~ N(+α, 1), that is, the temperature was a draw from a normal distribution with a mean of α and a standard deviation of 1.0. At island 2, the daily maximum temperature ~ N(-α, 1). 100-member ensembles of weather forecasts were generated by taking random draws from each island's climatology. 100,000 days of weather and ensemble forecasts were simulated, and we consider the event that the temperature was greater than 0. On island 1, both verification and ensemble ~ N(+α, 1) and were drawn independently. The same process was repeated for island 2, but verification and ensemble ~ N(-α, 1) .

Figure 1 synthesizes the forecasts scores' overestimate as a function of α when the Brier skill score is calculated by computing the climatology by eqs. 6-7, the ROC AUC is calculated using eqs. 10 – 11, and the ETS is calculated using eqs. 14 – 15. As α is increased, the forecast skill is progressively overestimated, even though the ensemble is always randomly drawn from each island's climatology.

What was the source of the overestimation of skill? For each of these scores, the computation no longer implicitly assumed that the climatological distribution was ~ N(+α, 1) **or** ~ N(-α, 1). Rather, it assumed that the climatological distribution was ~ 0.5 • N(+α, 1) +0.5 • N(-α, 1), a bimodal distribution when α is large. Meanwhile, the scores were computed consistent with the assumption that the forecast perfectly predicted which mode of the composite climatological distribution the verification lay in; when the forecasts were drawn from the positive mode N(+α, 1), the observed states were also drawn from the positive mode N(+α, 1), and when the forecasts were drawn from N(-α, 1), the observed state were drawn from N(-α, 1) as well. This illustrates that these scores can report false skill in situations where the climatologies differ among the samples used to populate the contingency tables; they credit a forecast with having skill merely if the climatologies of the individual samples are different from the climatology of the combined samples.

4. **Climatological forecasts of 850 hPa temperature**

We now demonstrate a simple example of false skill reported with real data. 0000 UTC 850 hPa temperature analyses were extracted from the 2.5° NCEP-NCAR reanalysis (Kalnay et al. 1996) at a set of 26x12 grid points covering the conterminous United States (US). Data was considered for the first ~ 2 months (60 days) of each year from 1979 to 2001, a mid-winter period when a grid point's climatological temperature distribution should be relatively stable, i.e., random samples from January 1 and February 28 (JF) can roughly be assumed to be sampled from the same underlying distribution. Let T denote the temperature at a grid point, and T ' denote the temperature anomaly from the mean. Two events were considered: (1) $T > 0$C, and (2) $T' > Q_{2/3}$, where $Q_{2/3}$ was the upper tercile of the climatological distribution, i.e., the temperature threshold defining the boundary between the lower two-thirds of the distribution and the upper third. $Q_{2/3}$ was specified uniquely for each grid point.

First we describe the method for generating contingency tables for the event $T > 0$C. For each of the first 60 days of the year and for each of the 23 years (1380 samples), the following cross-validated process (Wilks 1995) was performed at each grid point: (1) the analyzed temperature was extracted at that grid point, (2) the climatological probability of the event was determined using the other 22 years of data, (3) a 50-member ensemble was randomly drawn from the other 22 years of JF temperature samples at that grid point, (4) the ensemble was sorted, and (5) contingency tables were populated for that grid point. After all grid points were processed in this manner, average contingency tables for all of the grid points were also generated. To generate contingency tables for the ETS, the process was the same, but a single random sample from the 22 years of JF data was drawn rather than an ensemble.

When generating ROCs and ETSs for the event $T' > Q_{2/3}$, several additional steps were required. After step (1) above, the climatological mean for each date and location was determined and subtracted from the temperature, creating a database of temperature anomalies. The cross-validated climatological mean was estimated using a 30-day window centered on each day, using the remaining 22 years. Also, the terciles of the distribution were determined for each grid point.

a. T > 0 C

The climatological probabilities for this event varied from 0.005 in the north to 1.0 in the south. The mean climatological probability was 0.59 with a standard deviation of 0.36.

When a location-dependent reference climatology was used (eqs. 4-5), the BSS was -0.03. When the domain-averaged climatology was used (eqs. 6-7), the BSS reported a false skill of +0.52.

Figure 2a shows ROCs calculated from the individual grid point data; the ROC for every third grid point in the N-S and E-W directions are plotted. The ROCs exhibit sampling variability but lie close to the HR=FAR line. However, the ROC based on a contingency table summed up over all the grid points (Fig. 2b) diagnosed a very large amount of skill. Again, these were artifacts of the widely differing climatologies for the grid points, as in section 3.

Table 5 reports the ETS for this event. The ETS was calculated for each of the *m* locations using eq. 12 and then averaged. For some of these locations, the denominator of eq. 12 was zero and the ETS was undefined, so the average ETS reported in Table 1 was calculated excluding these locations, a tiny fraction of the number of grid points (see section 5 for an alternative method of calculation that does not exclude these locations). The ETS was also calculated using the summed contingency tables and eq. 14, excluding the same locations in calculating the table sums. As Table 5 shows, the averaged ETS was approximately zero, but the ETS from the table sums was 0.345, reporting a false positive skill because samples with different climatologies were mixed together into the same contingency table.

b. $T' > Q_{2/3}$

By evaluating the probability of exceeding a quantile of the distribution, the climatological probabilities have been rendered uniform across all grid points; the climatology probability is of course 1/3 for this event. Consequently, the BSS was the same for both, -0.03 (it was less than zero because the 50-member random draw from climatology only approximates the long-term sample climatology). With the ROC, whether we examined the average of scores at the grid points or computed the scores from contingency table sums, we found no skill (Fig. 3). Similarly, the ETSs (Table 5) reported the same lack of skill regardless of the how the ETS was computed.

5. **Equitable threat scores for numerical precipitation forecasts**

One of the important goals of the U. S. National Weather Service is to improve forecasts of precipitation. The ETS is one measure that is very commonly used to evaluate the skill of their deterministic forecasts. The most common approach is to estimate the ETS for fixed precipitation thresholds from a contingency table populated over many days or months and over a wide geographic region such as the conterminous US. We demonstrate here that the ETS calculated in this manner can drastically overestimate forecast skill.

To demonstrate this, a very large set of numerical forecasts was used, provided by the analog forecast technique discussed in Hamill et al. (2005). The details of the forecast methodology can be found in this reference but are not particularly important here. What is germane is that we produced a 25-year time series of gridded deterministic precipitation forecasts, all using the same model and forecast technique. These forecasts have characteristics similar to those of current operational forecasts. For this demonstration, we limit ourselves to considering the ETS of the mean of a 5-member ensemble of analog forecasts over the conterminous US for January and February from 1979 to 2003. Both the forecast and the verification data (from the North American Regional Reanalysis, Mesinger et al. 2005) are on a ~32 km grid. We consider the 5 mm precipitation threshold here.

Figure 4a illustrates the geographic dependence of the ETS on forecast location. This map is a very effective way of presenting information on the geographical dependence of threat score; skill was much larger in the southeast US and along the west coast than in the northern Great Plains. Perhaps a user requires the information to be condensed to a single number to facilitate comparison between two different forecast models. The ETS calculated from the contingency table sum using eq. 14 was approximately 0.41. However, examining Fig. 4a, it was apparent that the large majority of grid points had ETS much below 0.41, suggesting again that ETS was overestimated.

Unfortunately, there were many points in Fig. 4a evaluated with a zero ETS, points at which no forecasts of greater than 5 mm were issued during the period. This problem occurred for a much greater fraction of the grid points at higher precipitation thresholds (not shown). Clearly, we would prefer each contingency table to be populated with enough samples that the statistics are relatively stable. One possibility is to bin contingency tables together if they have similar climatologies. Assume we will bin together all grid points with climatologies discretized to the nearest percent. Define new contingency table elements $a_i^c, b_i^c, c_i^c, d_i^c, \quad i = 1, \ldots, 100$, where the ^{c} indicates a binning over climatologically similar forecasts. Since the sample event probability for the *j*th grid point is defined by $a_j + b_j$, the elements can be defined according to

$$a_i^c = \sum_{j=1}^{m} a_j \left| \left(\frac{i-1}{100} \le a_j + b_j < \frac{i}{100} \right) \right. \quad . \tag{16}$$

b_i^c, c_i^c and d_i^c are similarly defined. $a_i^c, b_i^c, c_i^c, d_i^c$ are then rescaled so their sum is again 1.0. Then the ETS can be defined for each bin in a manner similar to eqs. 12 – 13. Let $ETS^c(i)$ denote the ETS for the *i*th climatological bin, and let $f^c(i)$ denote the fraction of the grid points populating the *i*th bin. Then an overall threat score can be calculated according to

$$\overline{ETS}^{C} = \sum_{i=1}^{100} f^{c}\left(i\right) ETS^{c}\left(i\right) \qquad (17)$$

Figure 5 shows the $ETS^{c}(i)$ for each climatological bin, and the lower dashed line denotes $\overline{ETS}^{C}$ while the upper dashed line denotes the *ETS* calculated from the contingency table sums according to eq. 14. The solid line plots $f^{c}(i)$, normalized by the maximum frequency. The overestimate of the ETS is now readily apparent; $\overline{ETS}^{C}$ is much lower.

The ETS estimation technique in eqs. 16-17 has drawbacks. Notably, the climatological event probability was defined by the sample event probability $a_j + b_j$, a reasonable assumption with over two decades of winter forecast data. If the verification period is very short, then this sample probability may be a poor estimate of the long-term event probability; ideally a long, temporally and spatially dependent climatology should be used, if available. Nonetheless, these details should not obscure the main point, the dramatic overestimation of threat score that is possible when contingency table values are summed across grid points with different climatologies.

6. **Discussion**

The preceding examples have demonstrated that the Brier skill score, relative operating characteristic, and the equitable threat score must be used with care when verifying weather forecasts. Typically, the meteorological question being asked is something akin to “what is the Brier skill score of my forecast averaged over Europe?” The naïve approach for calculating the Brier skill score may be to compute it under the assumption that the climatology is invariant across the verification region. Similarly, when diagnosing the relative operating characteristic, or equitable threat score, a common step is to composite the forecast data into contingency tables that accumulate weather information across the domain. The preceding analysis showed that these diagnostics may falsely report more skill than truly exists in situations where the climatology differs across the domain. The more the climatology differs, the larger the falsely reported skill. By logical extension, false skill may also be reported if the verification samples span different seasons or even different times of the day with different climatologies but the data are still composited. All of these are examples of what is known as “Simpson’s paradox” in statistics.

Ideally, verification is done with large samples of forecasts, and it is relatively simple to calculate an appropriate spatially and/or temporally varying climatology as a reference. Of course, these ideal conditions are often not met; the climatological reference may be difficult to calculate, especially for intermittent, small-scale phenomena, and forecast sample size may be small. Clearly, the skills of the statistical meteorologist will be put to the test. The intent should at least be to design the verification method to minimize these problems, making at least relative inferences of skill (is model A more skillful than model B?) more trustworthy.

Below, we suggest some general guidelines that may be useful in adapting verification strategies to minimize this problem, as well as some considerations:

• In order to avoid reporting false skill, perhaps the researcher can alter his or her verification methodology. Alternative methodologies can be used that should not report false skill, such as: (1) if appropriate, analyze events where the climatological probabilities are the same throughout the sample (e.g., Buizza et al. 2003, Fig. 5, or Zhu et al. 2002). Section 4 demonstrated that, for example, relative operating characteristics, Brier skill scores, and equitable threat scores of climatological forecasts of *quantiles* of the 850 hPa temperature distribution did not report false positive skill. Regardless of whether the climatological means and variances are large or small, the fraction events classified as "yes" events are identical for different locations or times of the year (this methodology may less appropriate for variables like precipitation amount, since commonly we prefer to diagnose precipitation skill as a function of amount, and a chosen quantile could reflect very different amounts depending on the location). (2) If sample sizes are large enough, perform the calculations separately each for sub-sample with a different climatology, as in section 5. The data can then be displayed with an informative plot indicating the geographic or temporal variability of the skill scores. If the data must be summarized in some manner and the small sample size affects the manner of summarization, perhaps scores can be composited among grid points with similar climatologies, as demonstrated in section 5.

• The *specific details* regarding how the verification metrics are calculated should be fully described in journal articles and texts, since minor changes in the methodology can dramatically change the reported scores.

- Other scores such as the ranked probability skill score (Wilks 1995) can also falsely report positive skill, just as with the Brier skill score. Whatever the chosen verification metric, it is prudent to verify that climatological forecasts give the expected no-skill result before proceeding.
- Richardson (2001) demonstrated in a carefully controlled experiment that there was a theoretical equivalence between the Brier skill score and the integral of economic value assuming that users have a uniform distribution of cost-loss ratios between 0 and 1. One of the underlying assumptions was an invariant climatology across all samples. If this assumption is not met, then neither is this equivalence.

Acknowledgments

Dan Wilks and Matt Briggs (Cornell University), Craig Bishop (Navy/NRL), Beth Ebert (Australian BOM), Steve Mullen (U. Arizona), Simon Mason (IRI), Bob Glahn (NOAA/MDL), Neill Bowler and Ken Mylne (UK Met Office), Bill Gallus (Iowa State), Frederic Atger (Meteo France), Francois LaLaurette (EMCWF and Meteo France), Zoltan Toth (NCEP), and Jeff Whitaker (NOAA/CDC) are thanked for their discussions and comments on an early version of this manuscript. Three anonymous reviewers are thanked for their feedback. This research was supported by National Science Foundation grants ATM-0130154 and ATM-0205612.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918-932.

Appleton, D. R., French, J. M., and Vanderpump, M. P., 1996: Ignoring a covariate: an example of Simpson's paradox. *American Statistician*. 340-341.

Atger, F., 2003: Spatial and interannual variability of reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509-1523.

Bayler, G. M., R. M. Aune, and W. H. Raymond, 2000: NWP cloud initialization using GOES sounder data and improved modeling of nonprecipitating clouds. *Mon. Wea. Rev.*, **128**, 3911–3920.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78**, 1-3.

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the southwest monsoon. *Wea. Forecasting*, **17**, 1080–1100.

Buizza, R., and T. N. Palmer. 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518.

----------, T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **124**, 1935-1960.

----------, A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168-189.

----------, --------- , ---------, and --------, 2000a: Reply to comments by Wilson and by Juras. *Wea. Forecasting,* **15**, 367-369.

----------, J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 2000b: Current status and future development of the ECMWF ensemble prediction system. *Meteor. Appl*., **7**, 163-175.

----------, 2001: Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev*., **129**, 2329-2345.

----------, D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble prediction system and comparison with poor-man's ensembles. *Quart. J. Royal Meteor. Soc*., **129**, 1269-1288.

Chien, F.-C., Y.-H. Kuo, and M,-J. Yang, 2002: Precipitation forecast of MM5 in the Taiwan area during the 1998 Mei-yu season. *Wea. Forecasting*, **17**, 739–754.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev*., **129**, 2461–2480.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor*., **8**, 190-198.

Gallus, W. A., Jr., and M. Segal, 2001: Impact of improved initialization of mesoscale features on convective system rainfall in 10-km Eta simulations. *Wea. Forecasting*, **16**, 680–696.

------------, and -----------, 2004: Does increased predicted warm-season rainfall indicate enhanced likelihood of rain occurrence? *Wea. Forecasting*, **19**, 1127–1135.

Glahn, B., 2004: Discussion of verification concepts in “Forecast Verification: A Practitioner’s Guide in Atmospheric Science.” *Wea. Forecasting*, **19**, 769-775.

Göber, M., C. A. Wilson, S. F. Milton, and D. B. Stephenson, 2004: Fairplay in the verification of operational quantitative precipitation forecasts. *J. Hydrology*, **288**, 225-236.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

------------ , C. Snyder, D. P. Baumhefner, Z. Toth, and S. L. Mullen, 2000a: Ensemble forecasting in the short to medium range: report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653-2664.

------------ , ------------- , and R. E. Morss, 2000b: A comparison of probabilistic forecast from bred, singular vector and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835-1851.

------------ , J. S. Whitaker, and S. L. Mullen, 2005: Reforecasts, an important new data set for improving weather predictions. *Bull. Amer. Meteor. Soc.*, submitted. Available at http://www.cdc.noaa.gov/people/tom.hamill/reforecast_bams3.pdf .

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863-883.

Juras, J., 2000: Comments on “probabilistic predictions of precipitation using the ECMWF ensemble prediction system.” *Wea. Forecasting,* **15**, 365-366.

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-472.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150.

Kheshgi, H. S., and B. S. White, 2001: Testing distributed parameter hypotheses for the detection of climate change. *J. Climate*, **14**, 3464–3481.

Legg, T. P., K. R. Mylne. 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906.

Malinas, G., and Bigelow, J., 2004: Simpson's paradox. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed.. http://plato.stanford.edu/archives/spr2004/entries/paradox-simpson/

Marzban, C. 2004: The ROC curve and its area under it as performance measures. *Wea. Forecasting*, **19**, 1106-1114.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291-303.

---------- , 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Met. Mag.*, **37**, 75-81.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713-725.

-----------, and -----------, 2002: Areas beneath relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Quart. J. Royal Meteor. Soc.*, **128**, 2145-2166.

---------- , 2004: On using “climatology” as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev*., **132**, 1891-1895.

Mesinger, F., and coauthors, 2005: North American regional reanalysis. *Bull. Amer. Meteor. Soc*., submitted.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev*., **129**, 638–663.

-----------, and ----------, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor*., **10**, 155-156.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Royal Meteor. Soc*., **126**, 2013-2033.

Richardson, D. S. , 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc*., **126**, 649-667.

-----------, 2001a: Ensembles using multiple models and analyses. *Quart. J. Royal Meteor. Soc*., **127**, 1847-1864.

-----------, 2001b: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Royal Meteor. Soc*., **127**, 2473-2489.

Rogers, E., D. G. Deaven, and G. J. DiMego, 1995: The regional analysis system for the operational "early" Eta Model: original 80-km configuration and recent changes. *Wea. Forecasting*, **10**, 810-825.

-----------, and coauthors, 1996: Changes to the operational "early" Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting*, **11**, 391-413.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Simpson, E. H., 1951: The interpretation of interaction in contingency tables. *J. Royal. Stat. Soc.*, **13**, 238-241.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989. *Survey of common verification methods in meteorology*. Enviroment Canada Research Report 89-5, 114 pp. Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.

Stefanova, L., and T. N. Krishnamurti, 2002: Interpretation of seasonal climate forecasts using Brier skill score, the Florida State University superensemble and the AMIP-I data set. *J. Climate*, **15**, 537-544.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077-2107.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science*, **182**, 990-999.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Chapter 7 of "*Forecast Verification: A Practitioner's Guide in Atmospheric Science*." John Wiley and Sons, 254 pp.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks. 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev*., **129**, 729–747.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Cambridge Press. 547 pp.

-----------, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8**, 209-219.

Wilson, L. J., 2000: Comments on "probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Wea. Forecasting,* **15**, 361-364.

World Meteorological Organization, 1992: *Manual on the Global Data Processing System*, section III, Attachment II.7 and II.8, (revised in 2002). Available from http://www.wmo.int/web/www/DPS/Manual/WMO485.pdf.

Xu, M., D. J. Stensrud, J.-W. Bao, and T. T. Warner, 2001: Applications of the adjoint technique to short-range forecasting of mesoscale convective systems. *Mon. Wea. Rev*., **129**, 1395-1418.

Yang, Z., and R.W. Arritt, 2002: Tests of a perturbed physics ensemble approach for regional climate modeling. *J. Climate*, **15**, 2881–2896.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson and K. R. Mylne. 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc*., **83**, 73–83.

LIST OF TABLES

LIST OF FIGURES

University of Texahoma

	Admit	Deny
Female	120 (80%)	30 (20%)
Male	45 (90%)	5 (10%)

Table 1: Contingency table of numbers of students and admissions rates (in parentheses) into physical science programs at the University of Texahoma.

Texahoma State University

	Admit	Deny
Female	5 (10%)	45 (90%)
Male	50 (33%)	100 (66%)

Table 2: As in Table 1, but for Texahoma State University.

Both Universities

	Admit	Deny
Female	125 (62.5%)	75 (37.5%)
Male	95 (47.5%)	105 (52.5%)

Table 3: Contingency tables accumulated from both the University of Texahoma and Texahoma State University.

		Event forecast by *i*th member?	
		YES	NO
Event Observed?	YES	a_i (j) Mitigated loss (C+L_u)	b_i (j) Loss ($L = L_p + L_u$)
	NO	c_i (j) Cost (C)	d_i (j) No cost

Table 4: Contingency table for the *i*th of the *n* sorted members at the *j*th location, indicating the relative fraction of hits [a_i(j)], misses [b_i(j)], false alarms [c_i(j)], and correct rejections [d_i(j)]. The economic costs associated with each contingency are also shown and are discussed in the text.

	Event	
	$T > 0$	$T' > Q_{2/3}$
ETS (average of grid points)	-0.001	-0.002
ETS (contingency table sum)	0.345	-0.002

Table 5: Equitable threat scores for the events $T > 0$ and $T' > Q_{2/3}$, calculated as an average over all grid points with nonsingular ETS, and the ETS from the sum of contingency table elements at these grid points.

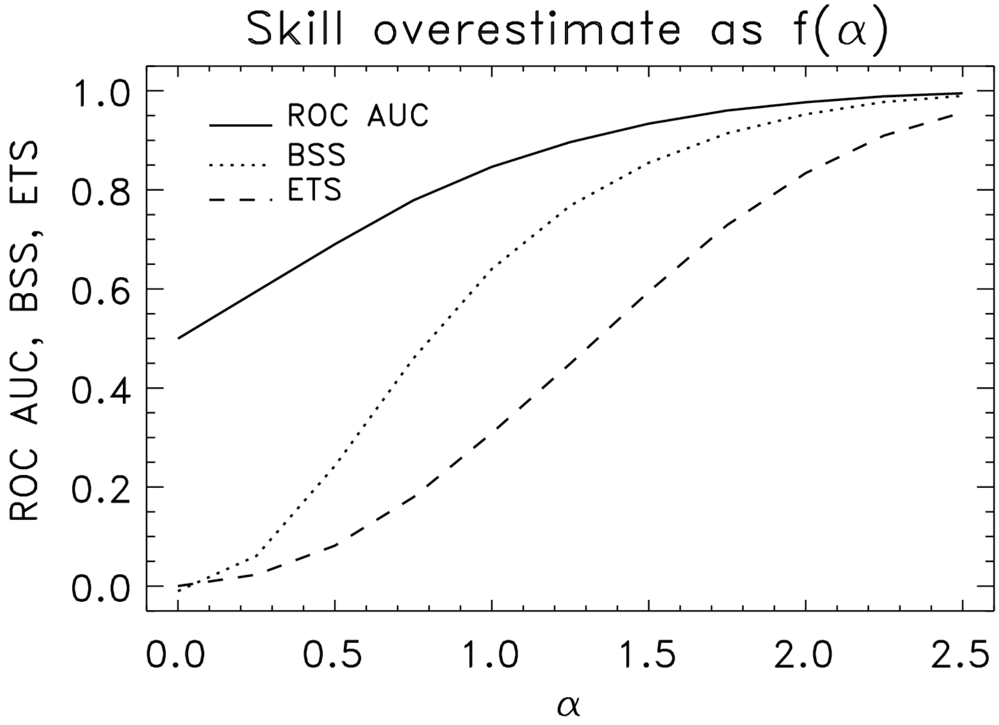

Figure 1: ROC AUC, BSS, and ETS as a function of the parameter α describing the difference in the means of the distributions between the two islands. Skill scores are calculated assuming a composite climatology.

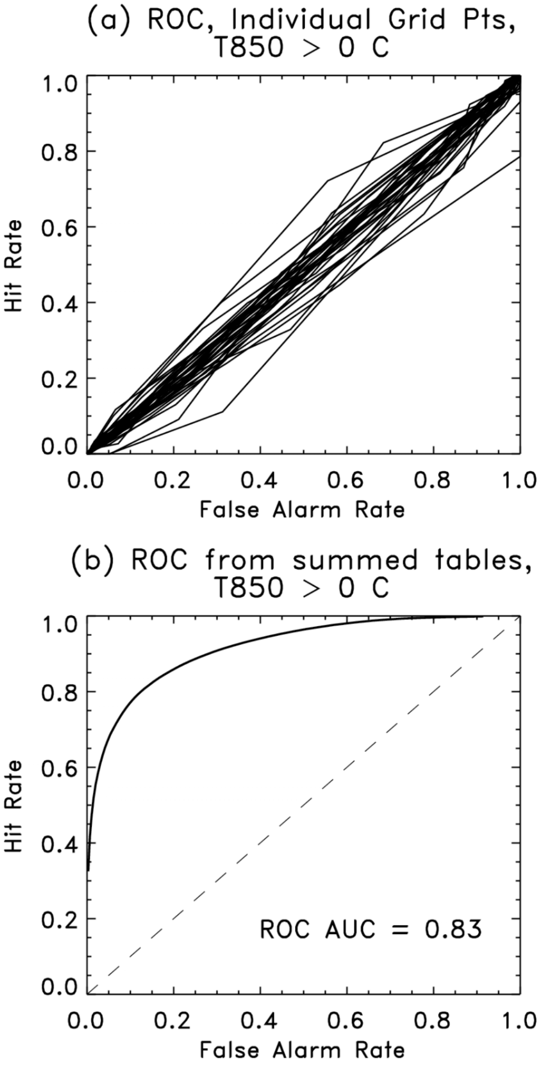

Figure 2: ROC for the event of 850 hPa temperature > 0 C using random draws from climatology using data from January-February 1979-2001. (a) ROC curves for selected individual locations around conterminous US, (b) ROC curve based on sum of contingency tables at individual grid points.

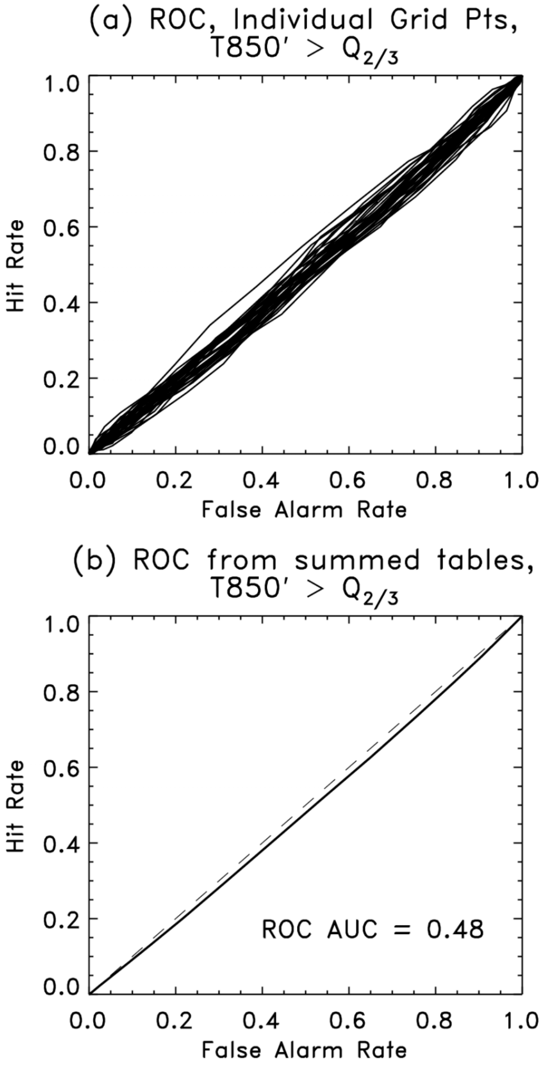

Figure 3: As in Fig. 2, but for the event of 850 hPa temperature anomaly is greater than the upper tercile of the climatological distribution.

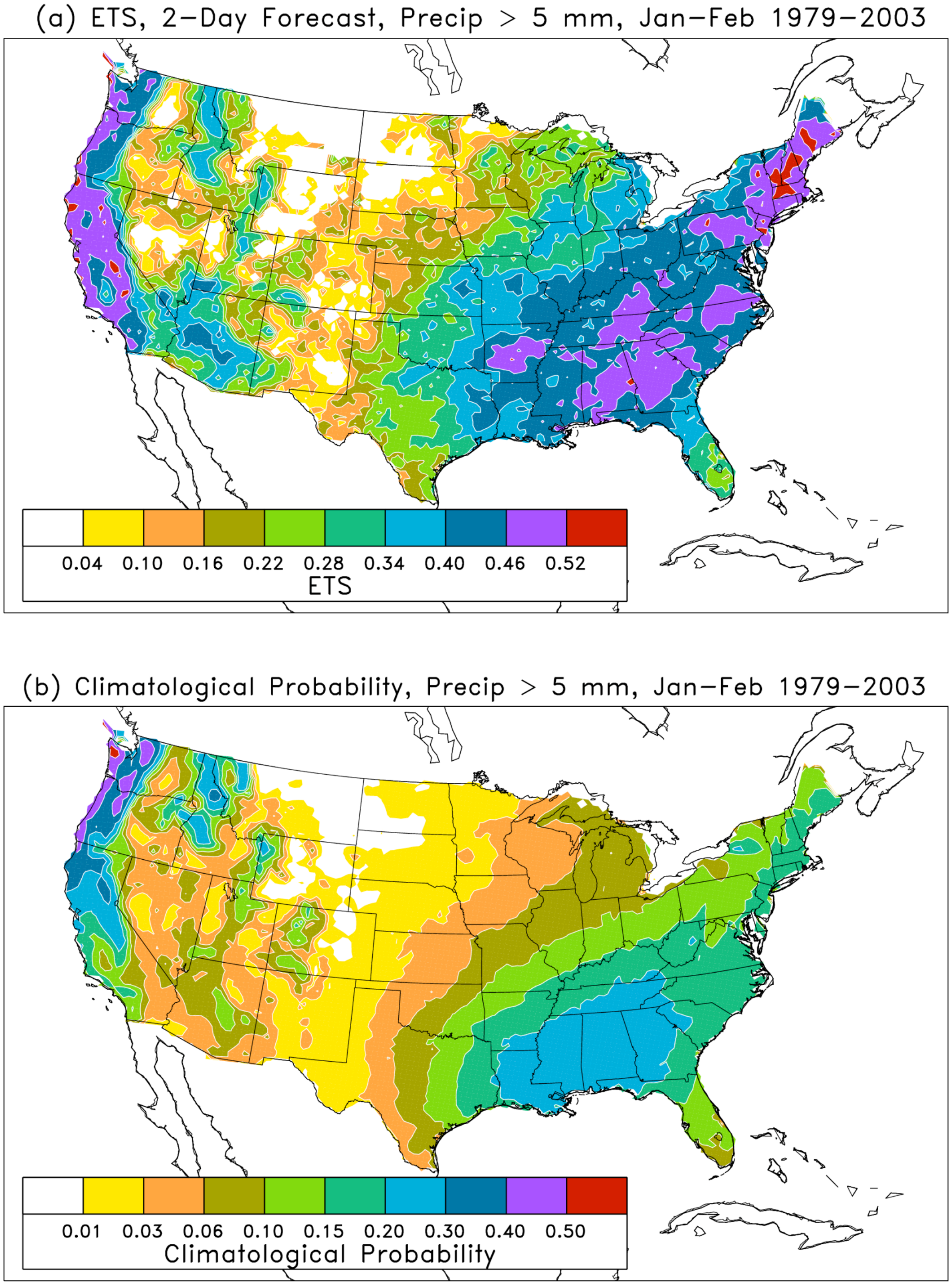
(a) ETS, 2-Day Forecast, Precip > 5 mm, Jan-Feb 1979-2003
0.04 0.10 0.16 0.22 0.28 0.34 0.40 0.46 0.52
ETS
(b) Climatological Probability, Precip > 5 mm, Jan-Feb 1979-2003
0.01 0.03 0.06 0.10 0.15 0.20 0.30 0.40 0.50
Climatological Probability

Figure 4: (a): ETS for 1-2 day 5 mm precipitation forecasts as a function of location, using Jan-Feb 1979-2003 forecast and observational data. (b) Climatological probability of precipitation greater than 5 mm for Jan-Feb.

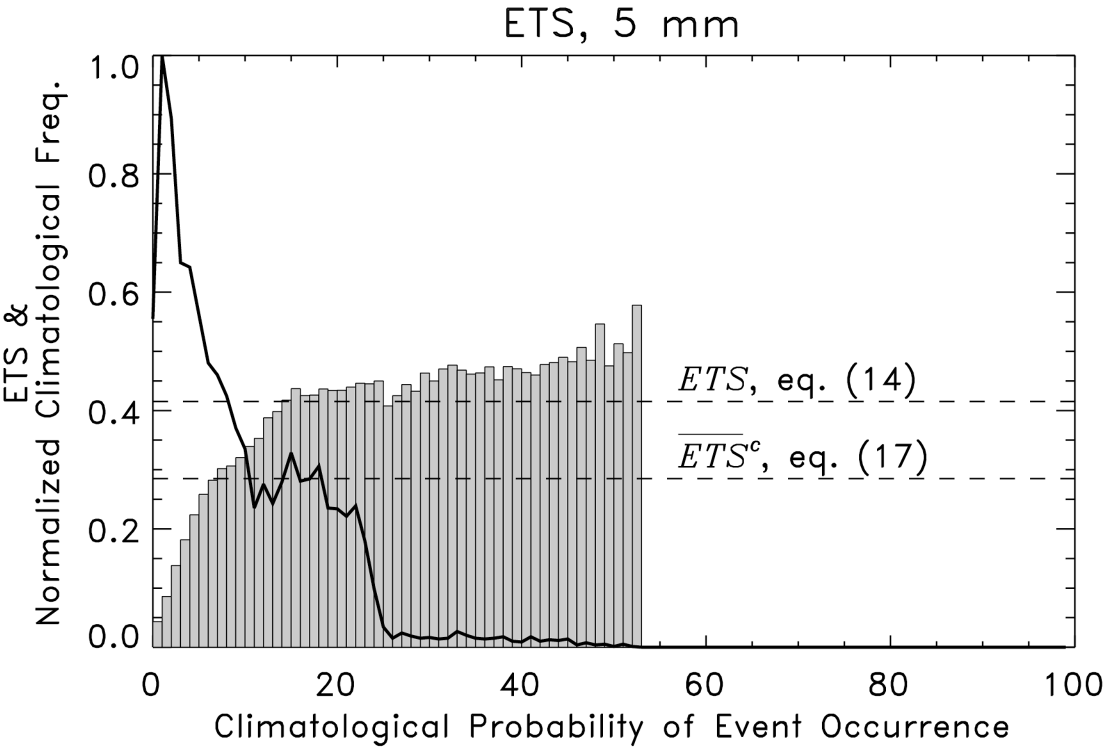

Figure 5: Equitable threat score (histogram) for the 5 mm threshold as a function of the climatological probability of event occurrence. The frequency of the climatological event probability is plotted as the solid line, normalized to a maximum value of 1.0. The dashed lines indicate the ETS calculated in two different ways.